Beryllium Health and Safety Committee
Data Reporting Task Force

White Paper

Draft, March 22, 2007

**Introduction**

On December 8, 1999, the Department of Energy (DOE) published Title 10 CFR 850 (hereafter referred to as the Rule) to establish a chronic beryllium disease prevention program (CBDPP) to:

- reduce the number of workers currently exposed to beryllium in the course of their work at DOE facilities managed by DOE or its contractors,
- minimize the levels of, and potential for, exposure to beryllium, and
- establish medical surveillance requirements to ensure early detection of the disease.

On January 4, 2001, DOE issued DOE G 440.1-7A, Implementation Guide for use with 10 CFR 850, Chronic Beryllium Disease Prevention Program, to assist line managers in meeting their responsibilities for implementing the CBDPP. That guide describes methods and techniques that DOE considers acceptable in complying with the Rule.

In 2005 a draft DOE Technical Standard "Management of Items and Areas Containing Low Levels of Beryllium" (SAFT 0103; hereafter referred to as the "TS") was circulated for comment (http://www.hss.energy.gov/NuclearSafety/techstds/tsdrafts/saft-0103.pdf). DOE technical standards are voluntary consensus standards developed when industry standards do not exist (see http://www.hss.energy.gov/NuclearSafety/techstds/index.html for more information). This TS is intended to supplement the Rule by describing best practices and lessons learned for managing items and areas that contain low levels of beryllium and are not included in the scope of the Rule, as well as determining if the Rule's housekeeping and release criteria are met.

Among the many comments on the draft TS was a suggestion that certain of the statistical comparisons described in the TS could be better implemented if "raw" data were to be reported by analytical laboratories. Exactly what is meant by "raw" in this White Paper is described in more detail below.

The Beryllium Health and Safety Committee (BHSC) formed a Sampling and Analysis Subcommittee (SAS) in 2003. The SAS established a working group on accreditation and reporting limits. By 2006 it had become evident that the issues extended to data reporting as a whole. The SAS proposed to the BHSC the formation of a Data Reporting Task Force (DRTF) to consider issues related to data reporting. The BHSC Board agreed, and requested that the DRTF generate a white paper, to be offered by the BHSC to potential interested parties such as the DOE beryllium policy office. It was noted that additional

products could include detailed guidance and potentially a journal article in the future. The SAS proposed that membership represent the affected disciplines (chemists, IH professionals, statisticians, and DOE-HQ policy). The BHSC Board decided that DRTF membership should come from DOE sites, since the focus would be on reporting in the context of the TS and the Rule. The DRTF came into existence in late 2006.

Teleconferences ensued, during which the following statement of purpose was developed: "The purpose of the Data Reporting Task Force is to harmonize perspectives among chemists, lab managers, statisticians, industrial hygienists, and policy personnel with respect to how laboratory reporting limits are determined, and how data at or below those limits are reported." A lack of a consensus on data reporting issues is one barrier to the adoption of a technical standard and the DRTF seeks to create consensus on these issues.

A face-to-face meeting of approximately half the members of the DRTF was held February 6 and 7, 2007. Many of the other members participated by telephone.

This White Paper summarizes issues discussed during the February 2007 meeting, and describes the path forward that was developed.

**Charter**

Since the DRTF is chartered by the BHSC, any recommendations developed by the DRTF will be delivered to the full BHSC for consideration, and then offered by the BHSC to potential interested parties such as the DOE beryllium policy office.

**Basic Terminology**

In this White Paper the terms "detection limit", "quantitation limit", and "reporting limit" are generic. They represent any of a large number of different methods and formulas that have been developed in the last 40 years or so to represent various aspects of an analytical method's ability to detect and measure the presence of a specified analyte in an environmental sample. One widely used detection limit is the U.S. EPA's "method detection limit", defined in 40 CFR 136 Appendix B (hereafter referred to as EPA MDL). The actual values of these limits, in practice, may be established by individual labs (and thus vary between labs), or may be client-specified.

In this White Paper the term "raw result" refers to a number, produced by an analytical measurement system, that is intended to estimate the concentration of the analyte in the sample. A raw result may be above or below any of the types of limits mentioned above, or even negative (< zero). Whether or not an analytical method can actually produce a negative number depends on the details of the method. Some do. Negative numbers are not interpreted to mean that there is a negative amount of the analyte in the sample, but rather that intrinsic analytical variation (including noise in the system) produces a scatter of values, and that scatter may include negative values, depending on the details of the

method. For example, the ICP-AES method when used for beryllium can produce negative values.

In this White Paper the term "censoring" refers to the practice of reporting an analytical result as "less than" a limit (any kind of limit), when the raw result is less than that limit. For example, if a raw result is 0.03, and the limit being used is 0.05, then a report of "<0.05" would be considered censoring. More specifically, this is left censoring. Right censoring, where a result as reported as "*>xx*" can occur as well; however, this is less relevant to the issues considered herein.

**The basic issue**

The draft TS proposed using statistical analyses of sets of samples to characterize potential beryllium contamination of facilities and equipment, and in particular, decide whether facilities or equipment may be released for uncontrolled use.

Surface wipe samples have been the main method used for such assessments and decisions. Results from relatively clean surfaces are often below laboratory reporting limits. Such censored ("less than") data is more difficult to interpret using conventional statistical methods than are detected results. This led to a proposal to report and use raw data to support statistically-based decisions on whether surface contamination exceeds action levels such as those specified in 10 CFR 850.

Thus, with regard to the primary issue of data reporting, there are two basic points of view:

- Using all raw data can improve statistical evaluations, and in particular allow release decisions to be made with statistical confidence with fewer samples than if censored, and should therefore be an acceptable way of reporting data (i.e., available upon request).

- Reporting raw data opens the door to misuse of the reported results, since raw data do not always represent detectable levels of Be. Raw data are at best highly uncertain, and in many cases are not "real numbers." Moreover, such reporting might threaten a laboratory's accreditation, be inconsistent with accepted technical practices, and create risk communication problems.

A second potential issue is whether or not the DRTF should make recommendations on what kind or kinds of limits should be used in the context of the laboratory accreditation requirement of the Rule. The topic of different types of limits, what they mean, and how they should be calculated, repeatedly arises during DRTF discussions.

**Discussion**

During the 2/2007 meeting the DRTF developed this "problem definition" list:

- We need a common set of terms for the limits used by laboratories analyzing for beryllium under 10CFR850, to promote consistency and minimize miscommunication.
- We need a consistent approach to calculating and using limits.
- We need to evaluate the No Censoring approach, and when and where it would be appropriate to use it
- We need to provide optimal approach(es) to reporting data based on identified data quality objectives (and thereby need to clarify DQO needs)
- We will need to provide education to end users regarding any recommendations we develop.

Decisions made based on the data need to be defensible. The way in which the data are reported must support the intended use. Therefore, it is possible that data may need to be reported differently, depending on the project or purpose for which the samples were collected.

There was some discussion at the 2/2007 meeting of what, exactly, are the regulatory drivers, and to what sampling situations and what types of decisions do they apply. For example, facilities (buildings) are different from equipment (computers, telephones, even an instrumentation trailer), and the decision process for them may be very different. It is easy to visualize a decision about a facility being based on summary statistics from multiple samples from the facility, but hardly practical to imagine a similar data analysis process for, say, half a dozen desktop computers. More likely, each will be assessed individually. The Rule itself establishes contamination control requirements for ongoing beryllium operations.

Section 10 CFR 850.31 "Release Criteria" establishes levels of cleanliness that should be achieved prior to release of contaminated equipment. Paragraph 850.31(b)(1) states "The removable contamination level of equipment or item surfaces does not exceed the higher of 0.2 μg/100 $cm^2$ or the concentration level of beryllium in soil at the point or release, whichever is greater." The rule does not specify how many samples should be used in making the determination, nor does it specify a metric (such as average or maximum) if more than one sample is used. The 0.2 μg/100 $cm^2$ surface loading criteria is based on experience indicating this level of cleanliness can be achieved with ordinary cleaning methods.

Due to the lack of more specific guidance, these criteria have been extended to the interpretation of surface wipe sampling for baseline surveys to determine whether legacy beryllium contamination from past operations may exist. The TS was drafted to create guidance more tailored to managing low levels of legacy contamination. The Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM) was used as a model for much of the TS. MARSSIM establishes sampling strategies and decision logic for the conduct of surveys for low levels of legacy radioactive contamination (see http://www.epa.gov/radiation/marssim/index.html.) However, reporting limits are not

addressed in MARSSIM as methods are usually able to measure background levels of radiation. MARSSIM assumes data will not be censored.

The decision criterion included in the TS is adopted from common Industrial Hygiene practice: a facility is declared "acceptable" if the 95%-95% Upper Tolerance Limit (UTL; the upper $95^{th}$ confidence limit for the $95^{th}$ percentile of the distribution of all possible measurements in the facility), is less than the 10 CFR 850 DOE Release Criterion (0.2 µg/100 $cm^2$). The following is a formal statement of the statistical test represented by the 95%-95% UTL procedure:

> Null Hypothesis; greater than 5% of surfaces in a survey unit exceed removable beryllium contamination limit; and
>
> Alternative Hypothesis; less than 5% of surfaces in a survey unit exceed removable beryllium contamination limit.

Although one would like to think that a preponderance of non-detections from a given facility should make it easier to declare a facility "acceptable", paradoxically it becomes more difficult to make that decision with higher proportions of censored data. This is because high proportions of censored data make the more efficient parametric statistical methods unavailable.

Note that the UTL formulation assumes that the facility is "not clean" unless and until a statistical test shows that (with a 5% "false clean" error rate) that it is "clean." This parallels, for example, EPA procedures for deciding that a waste stream is non-hazardous. One would hope that the cleaner the facility, the easier it is to show that it is clean using statistical methods, but a high proportion of censored data can make it more difficult.

[*The next paragraph is pretty much the statistician speaking, except that it does reflect something Paul said at one point about why he chose the criterion he did. Also, the next paragraph might be viewed as "them's fight'n words" by some … by all means toss it if it's controversial!]*

*-- as of 3/22/07 we have some variations on a discussion of the motivation for using statistically-based decision procedures rather than experienced-based decisions (i.e., expert opinion, professional judgment). First is the 3/16 original, followed by three alternatives, and then an integration –*

If a different decision criterion were to be employed, it is possible that the adverse effect of censoring on the data analysis would not be as great. However, the current statistically-based criterion was chosen, at least in part, to ensure a relatively objective decision based on a sufficient number of samples, rather than some sort of intuitive decision, along the lines of "enough of the data is far enough below the release criterion that we're confident the facility is ok".

-- nancy --
The choice of statistically-based criterion was made, at least in part, to foster uniform and objective decision based on a sufficient number of samples, rather than expert opinion or intuitive approaches that are not straight forward to defend or document. A standard procedure (followed by all users), documented by numeric values and mathematical calculations using recognized statistical procedures has provides for a strong legal basis for decisions.

-- melita –
If a different decision criterion were to be employed, it is possible that the adverse effect of censoring on the data analysis would not be as great. However, the current statistically-based criterion was chosen, at least in part, to ensure a relatively objective decision based on a sufficient number of samples, rather than some sort of intuitive decision, along the lines of "enough of the data is far enough below the release criterion that we're confident the facility is ok". One of the problems with this is the concern that raw data available for appropriate statistical analysis might be misused. Therefore, if raw data for statistical analysis is used, limits on the use of that data must be applied. [*note: I used the last two sentences in another section below*]

**Comment [m1]:** I am not sure where this statement really belongs but I think it needs to be emphasized somewhere.

-- paul –
If a different decision criterion were to be employed, it is possible that the adverse effect of censoring on the data analysis would not be as great. However, the current statistically-based criterion was chosen, at least in part, to ensure a relatively objective decision based on a sufficient number of samples. Statistically planned decision making is intended to control, to an acceptably low level of probability, the chance of contaminated "hot spots" occurring in an otherwise clean facility. It applies to decisions made from samples from a small percentage of the surfaces being characterized. Day-to-day risk management may require determining whether a specific piece of equipment or location is contaminated. These decisions are usually made from a single or a few samples intended to directly determine contamination levels of the surfaces that will be worked on. This is sometimes called diagnostic rather than statistically planned sampling and is not affected by censoring as long as the reporting limit is well below the regulatory limit.

-- end of alternatives --
-- and my integrated version, as of 3/22 --

If a different decision criterion were to be employed, it is possible that the adverse effect of censoring on the data analysis would not be as great. However, the current statistically-based criterion was chosen, at least in part, to ensure a relatively objective decision based on a sufficient number of samples, rather than expert opinion or intuitive approaches that are not straight forward to defend or document. A standard procedure (followed by all users), documented by numeric values and mathematical calculations using recognized statistical procedures provides for a strong legal basis for decisions. Statistically planned decision-making was included in the TS in order to control, to an acceptably low level of probability, the chance of incorrectly deciding that a facility is clean. It applies to

decisions made from samples from a small percentage of the surfaces being characterized.

-- end of alternatives and integrated version --

Day-to-day risk management may require determining whether a specific piece of equipment or location is contaminated. These decisions are usually made from a single or a few samples intended to directly determine contamination levels of the surfaces that will be worked on. This is sometimes called diagnostic rather than statistically planned sampling and is not affected by censoring as long as the reporting limit is well below the regulatory limit.

Current data reporting practices are driven by laboratory accreditation policies aimed at assuring results meet specified levels of accuracy. Standard practices for implementing these policies incorporate EPA methods mandated for analyses of waste water (method detection limits, MDLs, from 40CFR136; see the Terminology section above). These approaches have been used in EPA-based programs for many, many years, and are mandated for some work performed by environmental chemistry laboratories. Accreditation organizations view existing procedures as simple QA metrics and have therefore adopted them. For example, the AIHA requires the use of the 40CFR136 definition of MDL in its environmental lead program. This method has also been widely used for other metals for various reasons, including AIHA auditor requests, and to standardize MDL calculations with the lab. From the laboratory's point of view, if it must do a particular procedure for one program, then using the same procedure for other programs is easiest. There was an extensive discussion during the 2/2007 meeting of the history of EPA methods and their limitations.

The EPA has established a Federal Advisory Committee as a first step in court-ordered revision of its detection limit methods, which is controversial due to the potential regulatory ramifications. Overall, the DRTF concluded that EPA revisions to the MDL were unlikely to affect its work. More promising are other published methods for establishing reporting limits that can reduce censoring without being impractical due to complexity and be accepted by accrediting organizations as "equivalent" to EPA MDL methods. The DRTF charter does not include detection limit research or development.

Another external factor is a notice by the American Conference of Governmental Industrial Hygienist of its intent to lower its occupational exposure limit (OEL) for beryllium. Adoption of this OEL would make lower reporting limits for analyses of beryllium samples desirable. This may lead to the use of more sensitive instruments. Reductions in reporting limits are also possible through use of more complex calibration methods than those that incorporate the EPA MDL. This creates a reason for considering alternatives for establishing lower reporting limits that still meet accreditation policies for the accuracy of individual results.

Yet another external factor is the requirement of ISO 17025 to provide uncertainty information with reported data points. Addressing this requirement is among the options being considered by the DRTF.

Contamination characterization decisions at the facility level are most often made from aggregated data. In this setting, objectives can be met by using data in which individual results have lower levels of accuracy if they provide useful information on the distance between measured levels and the regulatory limits. Lower level raw data, which has greater relative uncertainty than results above a quantitation limit, still provide useful information on the distance between measured levels and regulatory limits. There is a concern, however, that low-level (below detection limit) raw data, if made available for appropriate statistical analysis, might be misused in other applications. Therefore, if raw data for statistical analysis is made available, it is expected that limits on the use of that data will be necessary.

Two types of intended use of analytical data need to be considered:

- Assessments based on the results of a single sample. These are typically exposure assessments for a single individual. For example, a single sample result is compared with an OEL.

- Assessments of a facility, item, or collection of items. These are typically comparisons of the levels in a facility with a release level, and are based on statistical analysis of data sets containing multiple samples. For example, a 95%-95% upper tolerance limit is compared with a release level.

The challenge for the DRTF will be to recommend methods that can accommodate these objectives while avoiding the potential for misuse or misinterpretation that accreditation policies are intended to prevent.

**Options**

The DRTF is considering two options for how to move forward. The first of these would be that the DRTF would not offer any new guidance. Laboratories performing analyses of surface samples for Be according to the TS and 10 CFR 850 would continue to report estimated concentrations only when they have good confidence in the number, and continue to report “less than” results as they have been.

A second option would be for the DRTF to offer guidance on the possible use of raw data for decisions based on data sets. Reporting of raw data would be negotiated between a project and its analytical laboratory; no lab would be forced to report such data, and the responsibility for its use (and protection against misuse) would be the responsibility of the project (not the lab). In order to ameliorate the accreditation issue, the official result, the one the Lab would be required to “stand behind”, would follow current practices (typically, censored at a reporting limit).

A third option would be for the DRTF to offer recommendations on the use of established standards, such as ISO 17025, to provide uncertainty information with reported data points. Some combination of the second and third options could be considered.

**The Path Forward**

The DRTF has identified four areas of activities that need to be performed, as follows:

- Writing – developing this white paper and offering guidance to interested parties. It could also include writing of other documents, such as voluntary consensus standards or journal articles, if that is deemed appropriate at a later time. Such activities would likely fall outside of the 12-18 month time window presently envisioned for the DRTF.
- Accreditation – ensuring that DRTF guidance is evaluated against existing AIHA accreditation requirements. If necessary, any needed changes to AIHA policy would be negotiated with AIHA, but in any case should remain consistent with ISO 17025.
- Technical issues – evaluating the technical issues before the DRTF for development of guidance as described above. Among these tasks include the following:
    - Development of a matrix listing the different needs for analytical data and the reporting requirements/options associated with those needs.
    - Evaluation of various types of reporting.
    - Evaluation of a performance-based "standard" for calculating reporting limits.
    - Evaluation of other options such as providing uncertainty intervals with the data.
- Education/risk communication – developing education and risk communication material to be presented to the beryllium community to aid in data reporting, statistical evaluations, and decision-making based on reported data.

The DRTF has agreed that the existing Accreditation Working Group (part of the Sampling and Analysis Subcommittee) should address the accreditation issues cited above. For the other activities, the DRTF has established sub-groups as follows:

- Writing Group: Don MacQueen (lead), Melecita Archuleta, Paul Wambach, David Weitzman
- Technical Issues Group: Tom Oatts (lead), Charles Davis, Nancy Grams, Burney Hook, Jim Robbins, Gary Whitney
- Education/Risk Communication Group: George Fulton, Don Harvey, Rohit Shah

The first two of these groups are currently active; the third group will become active at a later time when the other groups generate specific guidance. Until that time, members of the Education/Risk Communication group will participate with one of the two active groups.

The initial product of the DRTF is the present document, a draft version of the white paper, in time for review by the BHSC at its March 2007 meeting. A specific timeline for subsequent products has not yet been established.

**Additional Notes**

As an ancillary function, the DRTF will continue to follow the progress of the EPA Office of Water Federal Advisory Committee on Detection and Quantification (FACDQ), which will propose changes to 40 CFR 136 including the EPA's Method Detection Limit. The DRTF anticipates preparing comments during the public comment period on these changes. Such comments would be provided to interested BHSC members who may wish to use them to make comments.

The DRTF charter is for guidance on beryllium data reporting . However, similar issues exist with trace-level analyses in areas such as environmental monitoring. It may be that what the DRTF is able to accomplish for beryllium may be useful to other groups with similar issues in other contexts.

**Conclusion**

This white paper presents the reasons for the formation of the DRTF, a summary of the issues that the DRTF intends to address, its activities to date, and the path forward for resolution of the open issues. The benefits should include more consistent data reporting, better field decision-making, better understanding of the results by end users, and possibly a reduction (in some instances) in the number of samples required and associated sampling/analysis costs. Also, since data reporting issues were a key area of comments on the initial draft TS, guidance developed by the DRTF will be offered to the DOE beryllium policy office.

**Appendix – Detection limit concepts**

The DRTF believes it is important to understand detection limits, how they are developed, and what they mean, in order to develop recommendations for data reporting.

Most if not all detection limits in use today can be viewed as variations of the system of limits proposed by Lloyd Currie in Analytical Chemistry in 1968. These include:

- $L_C$: A level above which an instrument signal cannot credibly have come from a sample in which the analyte is not present. Such a signal is therefore considered to represent a "detection" of the analyte.
- $L_D$: A level that satisfies the following: when analyzing samples that are truly at or above $L_D$, the instrument signal is unlikely to be below $L_C$. That is, the analyte is highly likely to be detected (detected based on the $L_C$ criterion).

- $L_Q$: A level at which the estimated concentration in the sample has a low relative uncertainty (e.g., 10 to 20 percent).

In addition to the above is the concept of a reporting limit (or contract reporting limit), in which a laboratory and its client agree that if the estimated concentration is above the reporting limit, the laboratory will report the estimated concentration, otherwise it will report that the analyte was below the reporting limit.

**Bibliography**

Currie 1968 article

ISO 17025

Davis/Grams article

Draft DOE TS

40CFR136 Appendix B

EPA Revised Assessment of Detection and Quantitation Approaches, October 2004

MARSSIM

10CFR850

AIHA Policy Manual (LQAP)

AIHA Strategy for Assessing and Managing Exposures, Third Edition